Towards a Parallel Image Mining System

J. Fernández, R. Guerrero, N. Miranda, F. Piccoli *

Líneas Informática Gráfica y Paralelismo y Distribución del
Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Universidad Nacional de San Luis
Ejército de los Andes 950
5700 - San Luis - Argentina
e-mail: {*jmfer, rag, ncmiran, mpiccoli*}*@unsl.edu.ar*

Abstract

Images can reveal useful information to human users when are analyzed. The explosive growth in applying images as data in many fields of science, business, medicine, etc, demands greater processing power. With the advances in multimedia data acquisition and storage techniques, the need for automatically discovering knowledge from large image collections is becoming more and more relevant. Image mining, a relatively new and very promising field of investigation, tries to ease this problem proposing some solutions for the extraction of significant and potentially useful patterns from these tremendous data volume. This research field implies different stages, most of them demanding so many resources and computational time. The use of parallel computation is a good starting-point. Image mining process appears to be algorithmically complex requiring computing power levels that only parallel paradigms can provide in a timely way. As data sets involved are large, rapidly growing larger and images provide a natural source of parallelism, parallels computers could be organized to handle such big collection effectively. At this work we will examine the image mining problem with its computational cost, propose a possible global or local parallel solution and also identify some future research directions for image mining parallelism.

Keywords: Image mining, Image mining system, Parallel systems, Parallel techniques.

Resumen

El análisis de imágenes puede revelar información útil para los usuarios El significativo aumento del uso de imágenes en diferentes campos de la ciencia, medicina, negocios, etc., requiere de mayor poder de procesamiento. Con el avance en la adquisición de dato multimedial y de técnicas de almacenamiento, la necesidad de descubrir automaticamente conocimiento de grandes colecciones de imágenes aumenta. La minería de imágenes, area de investigación relativamente nueva y prometedora, trata de facilitar este trabajo proponiendo soluciones para la extracción de patrones significativos y potencialmente útiles a partir de grandes volumenes de datos. Comprende diferentes etapas demandantes de recursos y de tiempo computacional. El uso de computación paralela representa un buen punto de partida. El proceso de minería de imágenes parece ser algoritmicamente complejo, requiriendo niveles de poder computacional que solamente los paradigmas paralelos pueden proveer. Dado que involucra conjuntos de datos de rápido crecimiento y las imágenes representan una fuente natural de paralelismo, el paralelismo puede manejar semejante colección en forma efectiva. En este trabajo examinamos el problema de la minería de imágenes y su costo computacional, proponemos una possible solución global y local y definimos futuras extensiones para la minería de imágenes paralela.

Palabras claves: Minería de imágenes, Sistema de minería de imágenes, Sistemas paralelos, Técnicas de paralelismo.

*Grupo subvencionado por la UNSL y ANPCYT (Agencia Nac. para la Promoción de la Ciencia y Tec.)

1 INTRODUCTION

The tremendous growing of computerized information volume and variety has triggered the development of new data processing tools, World Wide Web technology and databases technologies that enable inferring useful knowledge from an important data bulk. As a result, it is necessary to support big collections of complex type information which includes complex objects data, spatial information or multimedia information. Many research works have focused on images and image mining.

The most general misinterpretation is that image mining only involves applying already existing data mining algorithms on images. Investigations in the area are usually pointed out into two main directions. The first one involves specific authority applications focusing on extracting most relevant image features, so they could be used in data mining [14][17][18]. The second direction applies to general applications, where the aim is discovering image patterns that might be useful in the understanding of existing interactions between human perception of the images at high level and image features at low level. Investigations in this direction try developments with major certainty of success in recovered images from a general purpose databases [13][20][24].

Human visual system has the ability to extract significant image relationships which are not represented in low-level primitive image features. Complex information and its use on specific applications leads to describe new association rules to information. The big challenge in image mining is extracting implicit knowledge, image data relationships, or other features not explicity stored in a pixel representation. As knowledge representation method, *patterns* have already been used by human being for simulating diverse cognitive process like intuition, intention and thinking. As long as the use of patterns can make the cognitive process more effective, they can be applied to describe the complexity and features of objects. Since the aim is to generate all significant patterns without any knowledge of the image content, diverse patterns types could be recognized: classification, description, correlation, temporal and spatial patterns.

Image mining deals with all aspects of large image databases including image storage, indexing schemes, and image retrieval, all concerning an image mining system [16]. Image databases containing raw image data as information, cannot be directly used for image mining purposes. Relational databases, traditionally used in data mining, do not satisfy this need; that is why other types of databases are defined like spatial, temporary, documentary and multimedia databases [26].

Figure 1 shows a general structure model for an image mining system. The system considers a specified sample of images as an input, whose image features are extracted to represent concisely the image content -Transformation and feature extraction phase-. Besides the relevance of this mining task, it is essential to consider invariance problem to some geometric transformations and robustness with respect to noise and other distortions while designing a feature extraction operator -Pre-processing phase-. After representing the image content, the *model description*

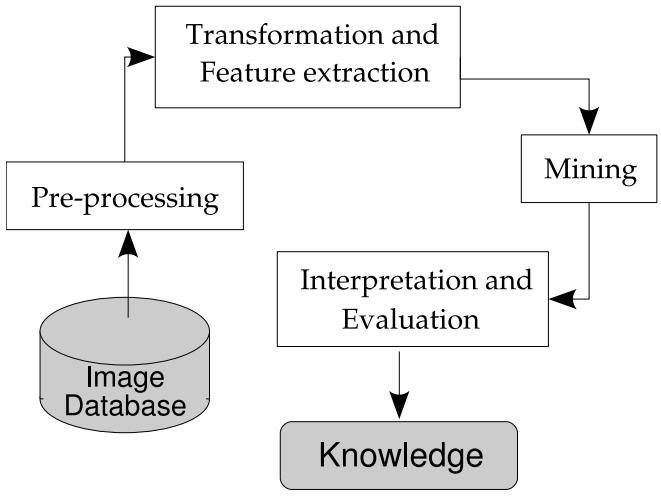

Figure 1: General Image Mining System

of a given image -the correct semantic image interpretation- is obtained. Mining results are obtained after matching the model description with its complementary *symbolic description*. The symbolic description might be just a feature or a set of features, a verbal description or phrase in order to identify a particular semantic.

The development of an image mining system is often a complex process since it implies joining different techniques ranging from data mining and pattern recognition up to image retrieval and indexing schemes. Besides, it is expected that a good image mining system provides users with an effective access into the image repository at the same time it recognizes data patterns and generates knowledge underneath image representation. Such system basically should assemble the following functions: image storage, image pre-processing, feature extraction, image indexing and retrieval and, pattern and knowledge discovery.

Image mining deals with the extraction of image patterns from a large collection of images, whereas the focus of computer vision and image processing is in understanding and/or extracting specific features from a single image. It might be thought that it is much related to content-based retrieval area, since both deals with large image collections. Nevertheless, image mining goes beyond the simple fact of recovering relevant images, the goal is the discovery of image patterns that are significant in a given collection of images. As a result, an image mining systems implies lots of tasks to be done in a regular time. Images provide a natural source of parallelism; so the use of parallelism in every or some mining tasks might be a good option to reduce the cost and overhead of the whole image mining process [1].

This works is structured as: the following section explains the differents difficulties and challenges involve in designing an image mining model. The section 3 explain the three main stages constituing a standard image mining system and their feasibility to be parallelize. Finally different parallel image mining models are proposed.

2 DIFFICULTIES AND CHALLENGES

Image mining deals with the study and development of new technologies that allow accomplishing this subject. A common mistake about image mining is identifying its scopes and limitations. Clearly it is different from computer vision and image processing areas. Moreover, the many knowledge discovery algorithms defined in the context of data mining are ill-suited for image mining. In image mining, there are many challengers still to overcome, some of them are:

- **Complexity of data:** To work with image and visual data is often to work with unstructured data, difficult to interpret and stored in a variety of different formats.

- **Scalability:** Image databases can easily reach hundreds of gigabytes and even terabytes in size. Scalable tools and algorithms for pre-processing and mining images that can manage such extremely large data in a reasonable time are yet to be developed. Massively parallel and high performance computing should help in this perspective for both image pre-processing and image mining.

- **Data inaccessibility:** Data adquisition and selection is fundamental in knowledge discovery process. The reasons for inaccessibility are multiple depending on the gathering means: sensors, satellites, among others. As incredible as it may seem, gathering images for research purpose or even industrial applications is not an easy task.

- **Privacy:** This has been an important issue with any data gathering and access. In some applications, image mining propels the problem of privacy a step further.
- **Minor support:** Image mining is relativity new, it relies heavily on fields such as vision and signal processing for data pre-processing and features extraction, fields which lack of adequate tool support, but in constant development.
- **Insufficient training:** Knowledge discovery and image mining tasks are related with many disciplines: artificial intelligence, databases, image and vision processing, high performance computing, visualization, etc. Interdisciplinary skills and work are required to process and cope with image.

To solve them in only one good application should be hard or impossible, but independent treatment of anyone could give notorious improvements to the whole image mining process.

3 PARALLELISM AND IMAGE MINING

Many issues of image mining can be optimized with different parallel techniques. Furthermore depending on tasks properties, different parallel paradigms could be applied in the same system. At a first glance, parallel applicant tasks will be: image storage, image processing, image indexing and retrieval and, pattern and knowledge discovery. In this section, the three main stages of an image mining system will be explained and then the feasibility of apply parallel paradigms at global and local level will be analized.

3.1 Processing Phase

Automatic image categorization involves experience on a real problem. The aim is to build a mining model using attributes extracted from and attached to the real problem, then evaluating the effectiveness of the model using new images. After the adquisition stage, the visual contents of the images in the database must be extracted and characterized by descriptive patterns - usually multidimensional feature vectors.

Orthogonal to challenges of developing specific image mining algorithms and models that operate on idiosyncrassy of images, one other major challenge for image mining is the pre-processing state previous to the extraction of relevant features. Generally, most of the images, if not all, are difficult to interpret, and a pre-processing phase is necessary to improve the quality of the images and make the feature extraction phase more reliable. The pre-processing state is arguably the most complex phase of the knowledge discovery process when dealing with images. If the pre-processing is well done, it can be decisive whether patterns could be discovered, or whether the discovered patterns could be interpreted at all. This phase often requires related expertise to computer vision, image processing, image interpretation, graphics and signal processing, domain knowledge or domain applications.

Pre-processing is always a necessity whenever the data to be mined is noisy, inconsistent or incomplete and it significantly improves the effectiveness of the data mining techniques. Nevertheless, a processing step is always done when applying any discovery technique for the extraction of relevant features. As a consequence, the process of building a mining model involves to split the processing phase into a pre-processing and extraction of visual features steps.

Figure 2 shows an overview of a categorization process. The first step is represented by the image acquisition and image enhancement, followed by feature extraction. The last one is the classification part, where different techniques for supervised learning are applied.

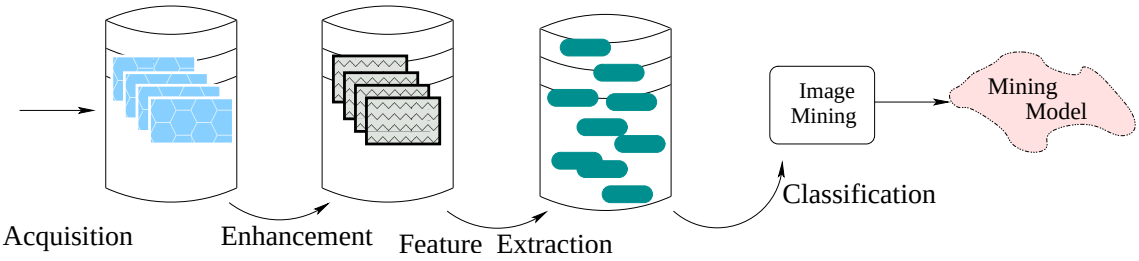

Figure 2: Image categorization process

The enhancement and feature extraction steps are usually refered as *Data Cleaning* and *Data Transformation* states that should be applied to the image collection. Data Cleaning is the process of cleaning the data by removing noise or other aspects that could mislead the actual mining process. Image enhancement helps in qualitative improvement of the image and can be done either in the spatial domain or in the frequency domain.

The most common techniques applied for data cleaning are the tipical image processing techniques like smoothing and sharpening filters. All these techniques could be combined with respect to a specific application.

On the other hand, data transformation implies to get an image content descriptor by means of its visual and semantic content. Visual content can be very general or domain specific. *General visual content* refers to color, texture, shape, spatial relationships, etc.; while *Domain specific visual content* is application dependent and may involve domain knowledge.

A good visual content descriptor should be invariant to any accidental variance introduced by the imaging process. A visual content descriptor can be either global or local. A global descriptor uses the visual features of the whole image, whereas a local descriptor uses the visual features of *regions* or *objects* to describe the image content. Moreover, as a previous step to obtain the local visual descriptors, an image is often divided into parts. The simplest way of dividing an image is to use a *partition*, which cuts the image into tiles of equal size and shape. A simple partition does not generate perceptually meaningful regions but is a way of representing the global features of the image at a finer resolution. A better method is to divide the image into homogenous regions according to some criterion using *region segmentation* algorithms that have been extensively investigated in computer vision. A more complex way of dividing an image, is to undertake a complete *object segmentation* to obtain semantically meaningful objects (like ball, car, horse).

Some widely used techniques for extracting color, texture, shape and spatial relationships from images are: Color Moments, Color Histograms, Color Coherence, Color Correlogram, Gabor Filter, Tamura features, Wavelet Transform, Moment Invariant, Turning Angles, among others [10][19].

Semantic content could be obtained by textual annotation or by complex inference procedures based on visual content. We will not discuss in detail this topic, trying to focus on our subject.

3.2 Mining Phase

In the nontrivial process of knowledge discovery in databases (*KDD*), data mining in general, and image mining in particular, have the aim of extracting implicit knowledge from data. They try to define valid, novel, potentially useful, and ultimately understandable patterns, relations or rules from them. These relations draw a *Predictive* or *Descriptive* model. With a predictive model is possible to estimate future or unknown values of interest, while with a descriptive model is possible to identify patterns which explain or summarize the analized data. Mining tasks depend in the model to be applied. *Classification* or *Regression* techniques define predictable models, while *Association Rule Mining* or *Clustering*, among others, define descriptive models. Image mining refers to a set of methods dedicated to the extraction of hidden knowledge from within an assortment of images. The early image miners have adopted existing machine learning and data mining techniques to mine for image information. Very few achievements have been realized and the approaches can be grouped in two classes. Those that discover patterns from:

- Images in large collections using the processed and extracted features within images;
- The image database using general descriptors.

While the applications vary from creating suitable models for image indexing to recognizing objects, categorizing images or image segments, the general tasks are similar and can be summarized as grouping images or features, either supervised or unsupervised, and associating image features.

The techniques frequently used include object recognition, image indexing and retrieval, image classification and clustering, association rules mining and neural network or a combination [8, 18].

3.3 Interpretation and Evaluation Phase

This task is a crucial one, it is tightly related with mining phase because it measures the quality from obtained patterns. Model preciseness can be secured by guarantying data independency between the training data set and testing data set.

Different evaluation techniques and measures can be applied. Evaluation measures could be objective or subjective. Which one of them would be used will depend on the mining taks to be done. The application context should be always considered when validating the obtained model [18, 13].

3.4 Global Parallel model

An image mining system(IMS) can be very computationally demanding due to the large amount of data to process, the response time required or the complexity of the involved image processing algorithms. Any parallel system requires dividing up the work so that processors can make useful progress toward a solution as fast as possible. The essential question is how to divide the labor.

There are three components to the work: computation, access to the data set, and communication among the processors [21]. These components are tightly related: dividing up the computation to make it faster creates more communication and often more data set accesses as well. Finding the best parallel algorithm requires carefully balance of the three named issues.

Parallelizing the image mining system showed at figure 1 involves to parallelize its three main areas: processing, mining and interpretation. Even though there exists a sequential line

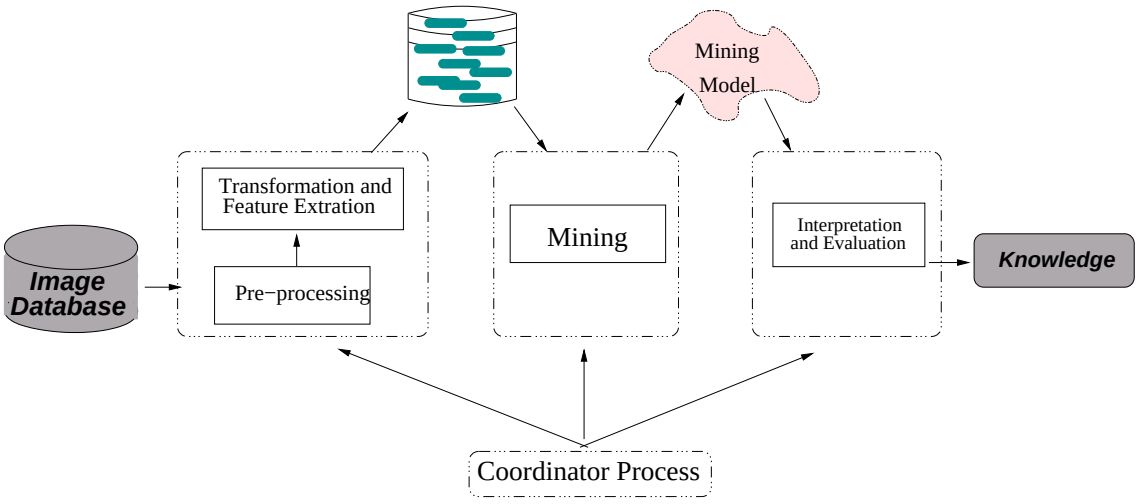

Figure 3: Global parallel arquitecture for the IMS

among them, it is a pseudo sequential line: after the first image set have been processed, the three named areas can be done in parallel.

Figure 3 shows the proposed global parallel arquitecture for the IMS. The model consists of four logical processing stages working in parallel: processing, mining, interpretation and coordination. The first three stages comes from the corresponding mining stages, and the last one from the derived management of the parallel model.

The coordinator process provides a GUI and task manager that directs the image mining process and is responsible for:

- At the starting of the mining process, it will coordinate the image mining tasks in a sequential way. First, the descriptor database generation through the processing phase, followed by the mining phase and finally the interpretation and evaluation phase.
- During the image mining process, it is responsible for the interaction with the image mining engine in terms of invoking, guiding and monitoring computations as well as visualization of the results.

Different parallel stage relationships are done by data sharing. Processing and mining stages share the feature database, and mining and interpretation stages share the mining model. Because reading and writing data structure accesses are simultaneous, synchonization mechanisms are required [25].

3.5 Local Parallel Model

At a refined level, each global parallel stage could be resolved in a parallel way. As a first attempt, we will focus only on the processing stage. This section sketchs three parallel levels concerning different parallel programming models and grains that could be accomplish collaboratively. The parallel alternatives are presented in an increasing complexity parallel order.

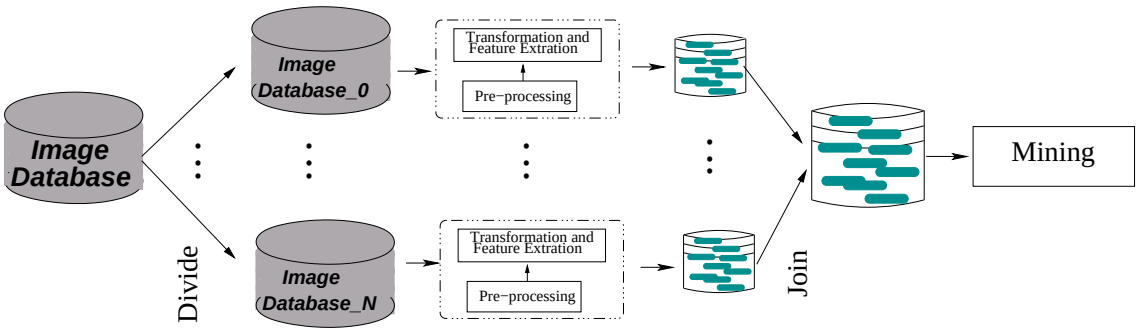

Figure 4: Level 1 Systems

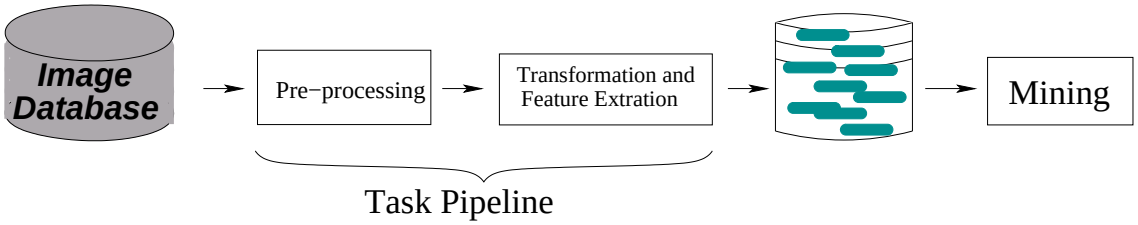

Figure 5: Level 2 Systems

3.5.1 Level 1: Embarrassingly Parallel

Applying the processing stage over the whole image set in the database could result in a time-expensive processing task. Besides this particularity, every internal data independence enables to draw a simple parallel model. The figure 4 shows the mentioned system architecture. N independent processes work on an image database partition, $DataBase\_i \quad (\forall i \quad 0 \leq i < N)$, making a feature data subset that will be joined to the whole working data set for the mining stage.

The parallel system has a coarse grain parallelism at data level following the $MDSP$ parallel programming model rules [2][15]. Moreover, as no particular effort is needed to segment the problem into a very large number of parallel tasks, and there is no essential dependency (or communication) between those parallel tasks, the problem is considered an embarrassingly parallel problem [25].

As each step can be computed independently from every other step, they could be made to run on a separate processor to achieve quicker results. An a-priori system performance estimation points out that it could be optimal or cuasi optimal.

3.5.2 Level 2: Parallelism into Processing Stage

At previous section only incoming system data independence was considered. At this section parallelism inside the processing stage will be take into account. Inside processing stage, as feature extraction step must be done after pre-processing, a pipelined processing is proposed, see figure 5 [25]. The pipeline has two well defined steps, the first one for image enhacement and the following for image feature extraction. As a consequence, a stream of images is passed through a succession of processes, each of which perform one task.

An interesting point to be considered in a pipelined parallel computation is the work de-

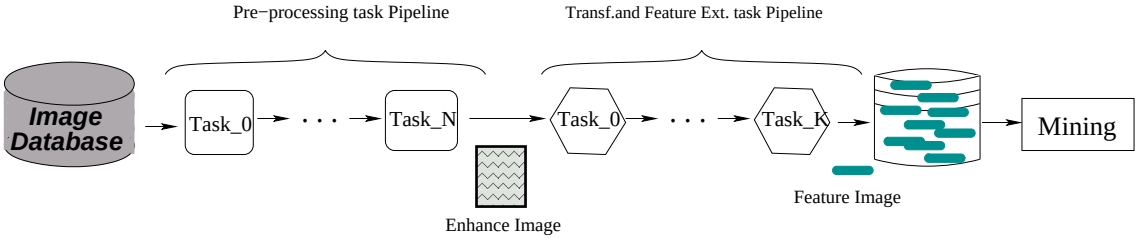

Figure 6: Refined Level 2 Systems

veloped into every step. When tasks workload is the same for every step, the pipeline gets the best performance. Either the enhancement or feature extraction steps address different processing task that should be done in an specific sequence. Dividing the enhancement or feature extraction steps (or both) into ordered substeps, could lead to take advantage from the inner step parallel characteristics and diminishing processing time problems due overwork into each of them. The system shown at figure 6 arises from the concepts stated. It can be observed a tasks pipeline where every task could be any classic processing task or a set of them, depending on the workload.

This last system has a finer parallel grain than the previous one where communications between tasks were increased.

3.5.3 Level 3: Parallelism Depending of Image Processing task

A sort of standard image processing tasks are commonly used at processing stage, like image smoothing, histogramming, 2-D FFT calculation, local area histogram equalization, local area, brightness and gain control, feature extraction, maximum likelihood classification, contextual statistical classification, image correlation (convolution, filtering), scene segmentation, clustering feature enhancement, rendering, etc. [6]. Many existing algorithmic implementions [3][5][7][9][11][12], could be done thru parallel solutions. Moreover, different techniques at different grain scale could be applied depending on the particular task; some of them are [4][22][23].

At this level any parallel model proposed not depends directly from the mining model itself, whereas it depends directly from any image processing task involved at the processing phase. As a consequence, any possible parallel model will be closely related to the specific image processing task to be done [10]; that is the reason because we do not suggest any model. The best solution could be to build a standard parallel image processing library that enables to make parallel processing at different combinations.

4 CONCLUSION

Integration of parallel tecniques into the image mining process was analized. It was considered from two points of view: global or local processing (each task into the global processing). Local analysis was focus on processing stage and three parallel image mining system models were proposed. They apply different parallel paradigm and grains: relations between comunication and computations.

At this moment, we are working on the parallel implementations of *level 2 system* as described in this paper. At the pipeline stage definition, the workload was considered. Moreover, the refined level 2 system is scalable and general. Scalability enables that any generic module

will be implemented, tested and assembled to the pipeline. Application domains other than image analysis will may also be benefited from the proposed methology. Generality enables to fit the pipeline states to the application context.

The current system has two stages implemented as a pipeline over a cluster of 15 nodes. The earlier results, besides they are few, are very promising.

REFERENCES

[1] H. Krawczyk A. Mazurkiewicz. A parallel environment for image data mining. In *Proceedings of the International Conference on Parallel Computing in Electrical Engineering (PARELEC'02)*, 2002.

[2] A.Grama, A. Gupta, G. Karypis, and V. Kumar. *Introduction to Parallel Computing.* Addison Wesley, 2003.

[3] D. Ballard and C. Brown. *Computer Vission.* Prentice Hall, Englewood Cliffs, 1982.

[4] J. Barbosa and J. Tavares A. Padilha. Parallel image processing system on a cluster of personal computers. *Lecture Notes In Computer Science*, pages 439 – 452, 2000.

[5] S. Beucher and F. Meyer. The morphological approach to segmentation: the watershed transformation. *Mathematical morphology in image processing*, pages 433–481, 1993.

[6] A. Choudhary and S. Ranka. Parallel processing for computer vision and image understanding. *IEEE Computer*, 25(2):7–9, 1992.

[7] J. Crespo, J. Serra, and R. Schafer. Theoretical aspects of morphological filters by reconstruction. *Signal Processing*, 2(47):201–225, 1995.

[8] C. Djeraba. *Multimedia Mining, A highway to Intelligent Multimedia Documents.* Kluwer Academic Publishers, 2003.

[9] C. Giardina and E. Dougherty. *Morphological Methods in Image and Signal Processing.* Prentice Hall, 1988.

[10] R. Gonzalez and R. Woods. *Digital Image Processing, 2nd Edition.* Prentice Hall, 2002.

[11] B. Jahne. *Digital Image Processing: Concepts, Algorithms, and Scientific Applications.* Springer Verlag, 1997.

[12] A. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall, 1989.

[13] Y. Keiji. Managing images: Generic image classification using visual knowledge on the web. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 167–176, November 2003.

[14] R. Kosala and H. Blockeel. Web mining research: a survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, June 2000.

[15] L.Yang and M. Guo. *High Performance Computing: Paradigm and Infrastructure.* Wiley-Interscience, 2006.

[16] R. Missaoui and R. Palenichka. Effective image and video mining: an overview of model-based approaches. In *MDM '05: Proceedings of the 6th international workshop on Multimedia data mining*, pages 43–52, New York, NY, USA, 2005. ACM Press.

[17] T. Mitchell, R. Hutchinson, M. Just, R.S. Niculescu, F. Pereira, and X. Wang. Classifying instantaneous cognitive states from fmri data. In *Proc. 2003 American Medical Informatics Association Annual Symposium*, pages 465–469, 2003.

[18] H. Orallo, R. Quintana, and F. Ramirez. *Introduccion a la Mineria de Datos*. Prentice Hall, 2004.

[19] J. Parker. *Algorithms for Image Processing and Computer Vision*. J. Wiley & Sons, 1997.

[20] A. Selim, K. Krzysztof, T. Carsten, and M. Giovanni. Interactive training of advanced classifiers for mining remote sensing image archives. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 773–782, 2004. (Industry/government track posters).

[21] D. Skillicorn. Strategies for parallel data mining. *IEEE Concurrency*, pages 26–35, 1999.

[22] W. Rapf M. ReinhardtL T. Braunl, S. Feyrer. *Parallel Image Processing*. Prentice Hall, Englewood Cliffs, Berlin Heidelberg, 2001.

[23] M. A. Vorontsov. Parallel image processing based on an evolution equation with anisotropic gain: integrated optoelectronic architectures. *Optical Society of America*, (16):1623–1637, 1999.

[24] Y. Wang, F. Makedon, J. Ford, L. Shen, and D. Goldin. Image and video digital libraries: Generating fuzzy semantic metadata describing spatial relations from images using the r-histogram. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 202–211, 2004.

[25] B. Wilkinson and M. Allen. *Parallel Programming: Tecniques and Aplications using Networked Workstations and Parallel Computers*. Prentice Hall, New Jersey, 1999.

[26] Ji Zhang, Wynne Hsu, and Mong Li Lee. Image mining: Trends and developments. *Journal of Intelligent Information Systems*, 19(1):7–23, 2002.